梁林林, 曲海成. 结合定位和局部判别性特征训练的细粒度图像分类[J]. 智能计算机与应用,2025,15(3):48-55. DOI:10.20169/j.issn.2095-2163.24081903

结合定位和局部判别性特征训练的细粒度图像分类

梁林林, 曲海成
(辽宁工程技术大学 软件学院, 辽宁 葫芦岛 125105)

摘 要: 为解决细粒度图像分类中不相关背景信息干扰,以及子类别差异特征难以提取等问题,提出了一种结合定位和局部判别性特征训练的细粒度图像分类方法(FLRDF)。FLRDF 是一个叠加的网络结构,首先利用 ResNet50 提取输入图片的全局特征,输出特征图,在训练过程中使用图像级别的注释,不使用物体边界框级别的注释、即弱监督目标定位;然后,对特征图进行判别性特征抑制处理,准确定位输入图像的前景目标,并按照原图尺寸对前景目标进行放大,再次送入 ResNet50 网络进行特征学习;最后,打乱前景目标的全局结构,使网络关注局部细节特征。整个网络训练过程,是在多分支损失函数的协同作用下共同约束网络的特征学习。经在 CUB-200-2011、Stanford Cars 和 FGVC-Aircraft 三个数据集的实验结果表明:本文方法均取得了不错的分类精度,分别为 88.1%、94.9%和 93.8%,证明本文算法优于 DCL、NTS-Net 等主流方法。

关键词: 细粒度图像分类; 弱监督目标定位; 局部判别性特征; 残差网络; 深度学习
中图分类号: TP181 **文献标志码**: A **文章编号**: 2095-2163(2025)03-0048-08

Fine-grained visual classification via foreground localization and region discriminative feature training

LIANG Linlin, QU Haicheng
(Software College, Liaoning Technical University, Huludao 125105, Liaoning, China)

Abstract: To address the interference of irrelevant background information and the difficulty in extracting subtle difference features among subcategories in fine-grained image classification, a fine-grained image classification method combining localization and local discriminative feature training (FLRDF) is proposed. FLRDF is a stacked network structure that firstly uses ResNet50 to extract global features from the input image, producing a feature map. During the training process, it employs image-level annotations without using object bounding box-level annotations, which is a form of weakly supervised object localization. The feature map is then processed with discriminative feature suppression to accurately locate the foreground targets in the input image. These foreground targets are enlarged to the original image size and fed back into the ResNet50 network for further feature learning. Finally, the global structure of the foreground targets is shuffled to make the network focus on local detail features. The entire network training process is constrained by a multi-branch loss function, which collaboratively guides the feature learning of the network. Experimental results on the CUB-200-2011, Stanford Cars, and FGVC-Aircraft datasets show that the proposed method achieves commendable classification accuracy rates of 88.1%, 94.9%, and 93.8% respectively, demonstrating that the algorithm outperforms mainstream methods such as DCL and NTS-Net.

Key words: fine-grained visual classification; weakly supervised object localization; region discriminative feature; ResNet; deep learning

0 引 言

细粒度图像分类是一种识别同一大类别下不同子类别的技术,如区分不同种类的鸟和不同型号的飞机。与通常的图像分类任务不同,细粒度图像分类更加注重识别具有判别性的局部位置信息,因此细粒度图像分类任务面临着艰巨的挑战。

细粒度图像由于是更精细的子类分类,所以主要难点是类间差异大、类内差异小,再加上光照、遮挡、形变等干扰,也增加了分类难度。

基金项目: 辽宁省高等学校基本科研项目(LJKMZ20220699)。
作者简介: 曲海成(1981—),男, 博士, 副教授,CCF 会员(46283M),主要研究方向:遥感影像高性能计算。
通信作者: 梁林林(1998—),女,助教,主要研究方向:图像与视觉信息计算。Email:767225128@qq.com。
收稿日期: 2024-08-19

深度学习细粒度图像分类方法，根据监督方式分为强监督训练和弱监督训练。这里展开分述如下。

强监督细粒度图像分类方法联合使用边界框和人工标注信息提升分类精度。Huang 等学者[1]提出的 Part-Stacked CNN 算法使用 2 个子网络，分别是定位网络和分类网络，并使用全卷积网络进行部分定位，将定位结果输入到分类网络；部件级网络通过共享层提取特征后计算关键点周围部件，再将部分网络及整体网络特征图合并后分类。Lin 等学者[2]提出的 Deep Lac 算法将定位、对齐和分类结合在一个网络里，提出了阀门联动函数（Value Linkage Function）进行算法中的反向传播，并形成了深度定位、对齐和分类系统。然而，强监督学习由于使用了额外的人工标注信息，增加了成本，并且容易出现标注错误导致的分类错误，限制了细粒度算法的性能。因此，近期的研究热点是在仅利用类别标签的情况下，采用弱监督的方式训练模型。

弱监督细粒度图像分类方法仅使用类别标注信息，不使用额外的标注。一些学者将视觉注意力机制应用到细粒度图像分类中。例如，Xiao 等学者[3]整合了 3 个注意力模型，将这 3 个注意力模型结合起来训练深度网络。然而，利用聚类算法得到的局部区域的准确性很有限。Zheng 等学者[4]提出了 MA-CNN 算法，该网络由卷积神经网络、通道分组层和分类子网络组成。首先，通过不同通道，对特征图的关注部分不同、其峰值也不同的特点，得到关键信息；其次，使用固定值进行裁剪，得到多个注意力图；最后得到预测值。然而，该算法忽略了重要语义部分和细粒度特征学习相互关联的事实。

虽然弱监督细粒度图像分类方法已经获得了一些的成果，但仍存在以下问题：

（1）输入图像的尺度不固定，变化较大，对特征学习有影响。

（2）若物体只占据图像的小部分，则很难准确提取到目标特征。

（3）细粒度图像具有类内差异大、类间差异小的特点，仅仅可以靠类之间的些许差别区分。

为此，本文提出了融合前景定位和局部判别性特征训练的细粒度图像分类方法。

1　相关工作

1.1　弱监督目标定位

弱监督目标定位仅使用图像标签定位输入图像中的目标对象，获取与真值框最贴切的边界框信息。其中，CAM 方法[5]是经典的弱监督目标定位方法，方法中的类激活映射图是由 CNN 的最后一个卷积层生成，可以为感兴趣的类别突出显示不同的对象区域。虽然基于 CAM 的方法简单有效，但也只能识别出对象的一小部分有区别的局部区域。为了提高 CAM 系列方法的定位性能，HaS[6]和 CutMix[7]采用了擦除输入图像的特征区域，迫使网络关注与目标相关部分的策略。与其不同的是，ACoL[8]和 ADL[9]删除了对应高响应特征的特征映射，并使用了多个对抗性训练的并行分类器提高目标定位性能。除此之外，SPOL[10]方法认为传统的融合方法很容易将底层特征掩盖在背景噪音中，所以 SPOL 通过对浅层特征和高层特征进行元素相乘，抑制背景噪音，并最大限度地将底层特性嵌入在浅层中。不同于以往的定位方法，本文提出了一个轻量级的弱监督目标定位，扩展目标激活响应的方法，实现精确的目标定位。

1.2　局部判别性特征学习

对于细粒度图像分类，学习局部区域的判别性特征是至关重要的，但这些特征在训练网络时往往会被忽略掉。因此，研究人员提出了很多加强图像局部区域特征提取的网络和方法。Ding 等学者[11]提出了 S3Ns 网络，通过稀疏注意图获得判别分支和互补分支，试图在保留图像上下文信息的同时获得局部特征信息。Yang 等学者[12]使用区域建议网络（RPN）找到更多的局部区域，通过置信度和非极大值抑制（NMS）选出最优的判别性局部区域，学习相应区域的特征。Chang 等学者[13]从损失函数的角度设计了一个多通道损失模型，以加强网络在全局和局部判别区域的表征能力。

对原始图像均匀分割后再打乱是细粒度图像分类的一种常见方法。如：DCL 方法[14]首先对输入图像进行均匀分割，然后破坏全局结构强调局部信息，最后重建图像来学习局部区域之间的语义相关性。PMG 方法[15]利用一个简单的拼图生成器，生成包含不同粒度信息的图像，让网络关注不同粒度的特征信息。

细节学习是针对定位后的前景图像进行随机打乱，一是可以减少无关的背景噪音，二是不会引入额外的参数开销。本文是基于破坏输入图像的整体结构、让网络关注不同的局部判别性特征的思想，提出了不同于 DCL 方法的细节学习。

2 本文方法

为了解决网络在提取特征时受到背景噪音的影响,以及无法学习多样化的局部区域特征等问题,本文提出了结合定位和局部判别性特征训练的细粒度图像分类方法,整体框架如图 1 所示。该框架主要由 3 部分组成:

(1)前景定位模块,定位输入图像的前景目标,根据定位结果重新放大到原图尺寸,消除背景噪声对特征提取的干扰。

(2)细节学习,打乱输入图像的全局结构,重组局部区域,提取局部细微且重要的特征信息。

(3)设计多分支损失函数,共同约束网络学习特征的能力。

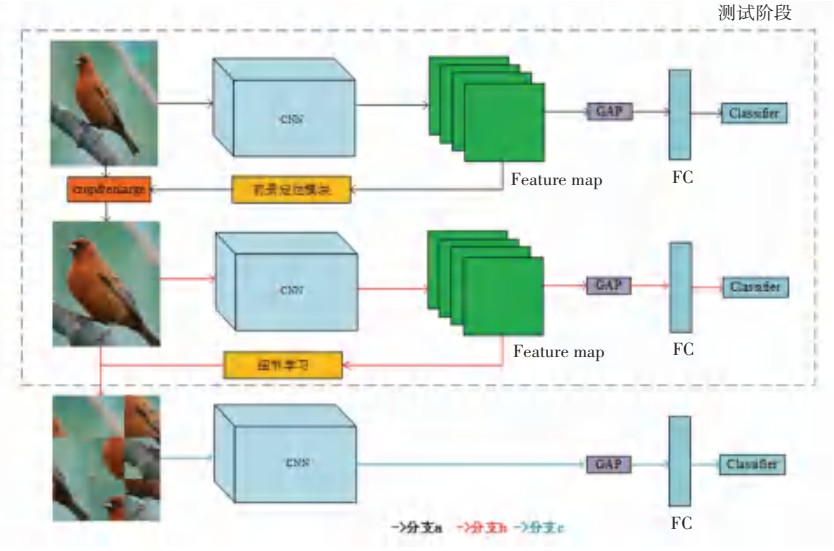


图 1 本文方法模型结构

Fig. 1 The proposed model architecture

结合前景定位和局部判别性特征训练的细粒度图像分类方法是三分支网络结构,使用 ResNet50 作为特征提取网络,共享 ResNet50 网络的全部参数信息。首先,将图片送入预训练的 ResNet50 提取输入图像的整体特征(图 1 分支 a),前景定位模块需要借助分支 a 中原始图像的特征映射来获取对象的边界框信息,剔除边界框以外的无关背景信息,同时上采样到原图尺寸,生成前景图像,送入图 1 中分支 b;然后,通过 ResNet50 网络学习前景目标的全局结构特征;最后,通过细节学习破坏前景图像的结构特征,学习前景图像的局部细微且多样化特征。整个过程以多分支损失函数优化模型收敛性。训练阶段是由三分支共同完成,测试阶段由分支 a 和分支 b 完成。

2.1 前景定位

为了得到精确的目标定位图,本节提出了前景定位模块(Foreground Localization module, FL)。利用 CAM 方法实现目标定位的主要问题是判别性特征区域的提取特征不充分,特征较分散。为了解决上述问题,前景定位方法主要将对判别性特征的关注扩展到相邻的非判别性特征区域。具体来讲,前景定位方法抑制了网络对判别性特征的关注,却提高了对非判别性区域的关注,从而提高了目标定位的准确性。前景定位结构如图 2 所示。

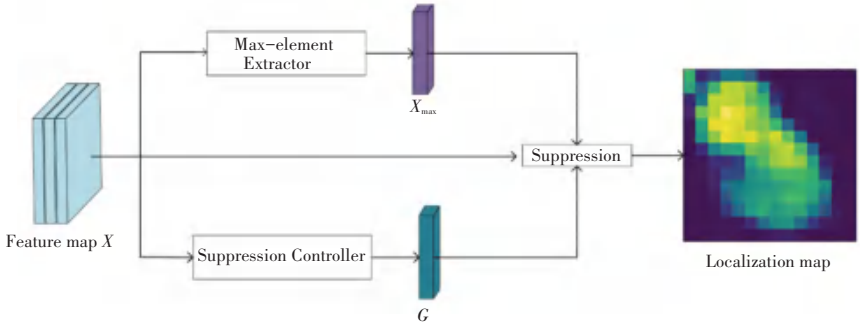


图 2 前景定位结构图

Fig. 2 Foreground localization architecture

假设 $X \in R^{H\times W\times K}$ 表示输入图像的最后一个卷积特征图，其中 K 表示通道数，空间大小为 $H \times W$。前景定位方法由峰值提取器、抑制控制器和抑制器三部分组成。

峰值提取器使用全局最大池化，从特征图 X 中提取 K 个最大元素，输出记为 $X_{\max} \in R^{1\times 1\times K}$。在此基础上，将这 K 个极大值元素作为判别区域的判据，并将其作为抑制的起点。

抑制控制器决定了对判别区域的抑制程度。运行生成了 $G \in [0,1]^{1\times 1\times K}$，$G$ 中的每一个第 k 个控制值决定了 X 中对应第 k 个最大元素的抑制量。如果抑制能力过强，则会削弱判别特征提取能力。可学习控制器自适应平衡分类网络的判别特征提取能力和抑制能力。形式上，可学习控制器的输出表示为：

$$G = \sigma(f(GAP(X);\theta)) \tag{1}$$

其中，f 表示全连通层；θ 表示控制器的可学习参数；$\sigma(\cdot)$ 表示 Sigmoid 函数。由于 θ 是与分类目标一起训练的，因此具有可学习控制的前景定位自适应抑制判别区域，使判别特征提取能力不受太大影响。

抑制器利用 K 个最大元素和 K 个控制值对判别区进行抑制。具体来讲，将 $X_{\max}$ 和 G 的元素点乘，重新定义为 X 的上界，记为 $\tau = X_{\max} \cdot G$，其中，$\tau \in R^{1\times 1\times K}$ 将 X 中高于该最大值的区域视为要抑制的判别区域；然后将 τ 扩展到与 X 相同的形状后，对 X 和 τ 进行取小操作，用来抑制高响应区域。例如：第 k 个控制值为 0.7，如果元素值没有超过第 k 个最大值的 70%时，X^k 被抑制。通过这种方法，抑制器在判别性特征区域和非判别性特征区域间架起了桥梁。

2.2　细节学习

细节学习是细粒度图像分类中重要部分之一，由于细粒度图像具有类内差异大、类间差异小的特点，局部信息要比整体结构更重要。所以，假定图像的局部区域被“打乱”，神经网络需要从判别性信息中学习分类细节。基于这一思想，本节提出了细节学习方法。细节学习就是通过破坏输入图像的全局结构，用于训练网络关注局部细节特征，同时辅以定位后的前景图像，强化网络训练的过程。为了防止破坏图像结构引入的噪音模式对网络学习的影响，细节学习是对定位后的前景图像进行破坏。因此，即减少了无关的噪音信息，又不会引入额外的参数开销。

细节学习如图 3 所示。细节学习是通过打乱整体图像的空间布局，使其关注局部细节特征。给定一个输入图像 X，首先将其平均分成 $n \times n$ 个子区域，然后对各个子区域进行随机打乱，合并成一张新的图像。子区域的大小由超参数 n 控制。分类网络需要找到判别区域并学习类别之间的细微差异。

（a）定位图像　　（b）打乱定位图像　　（c）细节学习

图 3　细节学习

Fig. 3　Delicate learning

2.3　损失函数

本文算法使用 3 个分支提取特征（完整算法见图 1）。其中，分支 a 提取整体特征；分支 b 根据原始图像映射得到边界框，缩放边界框放到输入图像上，实现前景定位；分支 c 打破输入图像的全局结构，重组局部区域，使网络学习到目标的局部特征。

3 个分支使用交叉熵函数作为分类损失，表示形式为：

$$L_{\text{raw}} = -\ln(P_r(c)) \tag{2}$$

$$L_{\text{object}} = -\ln(P_o(c)) \tag{3}$$

$$L_{\text{jig}} = -\ln(P_d(c)) \tag{4}$$

其中，c 表示输入图片的类别标签，P_r、P_o、P_d 分别表示 3 个分支中最后一个 Softmax 层输出的类别概率。总损失函数表示为：

$$L_{\text{total}} = \alpha L_{\text{raw}} + \beta L_{\text{object}} + \gamma L_{\text{jig}} \tag{5}$$

总损失函数是 3 个分支损失函数之和，用以优化模型在反向传播时的性能。3 个分支损失函数的协同合作，可以加快网络模型的收敛速度和提高模型的特征学习能力。在测试阶段，仅有图 1 中分支 a 和分支 b 完成分类任务，模型完成对原始图像输出粗粒度分类预测概率和前景图像输出细粒度分类预测概率后，最终分类结果取粗粒度预测和细粒度预测的平均值。

3　实验结果与分析

3.1　实验细节

3.1.1　数据集与预处理

为了验证本文算法的性能，分别在 3 个细粒度图像数据集 CUB－200－2011、FGVC－Aircraft 和 Stanford Cars[16] 上进行了验证，数据集的详细信息见表 1。

表 1 数据集信息

Table 1 Dataset information

数据集	对象	类	训练集	测试集
CUB-200-2011	bird	200	5 994	5 794
FGVC-Aircraft	aircraft	100	6 667	3 333
Stanford Cars	car	196	8 144	8 041

由于每个类别的图像较少,所以在训练前进行了数据增强。数据增强方式如图 4 所示。图 4 中,图片的水平翻转和垂直翻转均以概率 0.5 进行数据扩充。

(a) 原图

(b) 缩放

(c) 水平翻转

(d) 垂直翻转

(e) 亮度

图 4 数据增强

Fig. 4 Data augmentation

3.1.2 实验环境及参数设置

本文使用预训练的 ResNet50 作为基础网络,所有输入图片通过双线性插值调整到 448×448 大小。训练阶段对图片进行水平和垂直翻转,测试阶段仅对图片做归一化操作。

在参数设置上采用随机梯度下降法(Stochastic Gradient Descent,SGD)优化模型,动量为 0.9,权重衰减为 0.000 1, *epoch* 为 200, *batch* 为 6;初始学习率为 0.001,每经过 60 次 *epoch* 学习率乘上 0.1[17]。损失函数中将 α 设置 1, β 设置为 1, γ 设置为 0.1。

实验设备如下:实验环境为 Ubuntu 18.04.5, GeForce RTX 1080 Ti,运行内存为 128 GB,使用 1 个显卡进行训练。模型训练平台为基于开源深度学习框架 PyTorch,版本为 PyTorch 1.2.0,Python 版本为 Python 3.7。

3.2 评价指标

本文选择分类准确率作为评估标准,且为了证明前景定位方法的有效性,还引入了 *IoU* 作为检测边界框的评价指标。分类准确率和 *IoU* 分别为:

$$Accuracy = \frac{R_a}{R} \tag{6}$$

$$IoU = \frac{\text{area of overlap}}{\text{area of union}} \tag{7}$$

其中, R 表示测试集的图片数量, R_a 表示测试实验中正确分类的样本数量。

3.3 实验结果对比

3.3.1 消融实验

为了进一步验证模型中各个部分的有效性,在 CUB-200-2011、Stanford Cars 和 FGVC-Aircraft 数据集上进行了消融实验,并分析了不同方法消融研究的定量结果。本文提出的融合前景定位和局部判别性特征学习方法包括前景定位组件和细节学习组件。实验结果见表 2。

表 2 消融实验

Table 2 Ablation experiment

方法	Cub	Cars	Aircraft
ResNet50	85.5	92.7	90.3
前景定位	87.6	94.4	92.9
细节学习	80.5	91.4	89.8
本文方法	88.1	94.9	93.8

表 2 定量分析了每个方法组件的影响,单一的基于前景定位或细节学习的方法与一些 STOA 方法相比,对细粒度图像分类数据集的分类精度仍略显不足。但是,当前景定位方法和细节学习相结合时,其性能超过了一些 SOTA 方法,其中包括 PMG、

DCL、NTS-Net、MC-Loss、DFL-CNN 和 Cross-X。弱监督目标定位的引入，有效地使 ResNet50 从特征图中排除噪声，纯化特征提取；细节学习破坏定位图像的整体结构，让 ResNet50 从特征图中提取局部判别信息，从而提高了分类精度。

本文以 *IoU* 大于真值框的一半作为评价指标进行实验分析，实验结果见表 3。

表 3　前景目标定位准确率

Table 3　The accuracy of foreground object localization

方法	Backbone	Cub
CAM	ResNet50-GAP	65.7
SCDA[18]	ResNet50	76.8
ADL	ResNet50	62.3
前景定位（本文）	ResNet50	82.4

表 3 展示了不同的弱监督目标定位方法的实验结果，从表 3 的实验结果可知，本文提出的弱监督目标定位方法定位效果优于上述方法。

3.3.2　参数分析

为了确定超参数 n 对细节学习的影响，分别在 CUB-200-2011、FGVC-Aircraft 和 Stanford Cars[16] 数据集上进行验证，实验结果见表 4。表 4 展示了在输入图像大小为 448×448 的情况下，超参数 n 对 3 个数据集的分类精度。

表 4　参数分析

Table 4　Parameters analysis

n	Cub	Cars	Aircraft
1	87.63	94.37	92.93
2	87.25	94.11	**93.80**
3	**88.10**	**94.85**	93.58
4	87.33	94.44	93.14
5	86.20	93.90	92.68

随着超参数 n 的增加，分类精度先提升后降低。在 CUB-200-2011 和 Stanford Cars 数据集中，当 $n=3$ 时分类精度最高。对于 FGVC-Aircraft 数据集，在 $n=2$ 时取得了最好的分类精度。一方面，如果将 n 设置一个较大的数，细节学习从区域中学习到的视觉模式将受到限制，本文方法将难以收敛；另一方面，将 n 设置一个较小的数，细节学习方法将难以发挥优势。当 $n=1$ 时，此时的性能相当于前景定位基准线。

3.3.3　对比实验

本文算法在 CUB-200-2011、Stanford Cars 和 FGVC-Aircraft 数据集上进行充分实验，并与一些先进的方法进行对比，实验结果见表 5。表 5 中的实验数据表明，与其他主流的弱监督细粒度分类方法相比，本文方法在 3 个数据集上都取得了出色的分类成绩。其中，在 FGVC-Aircraft 数据集的精度比 PMG 方法提高了 0.4%，但是在 CUB-200-2011 和 Stanford Cars 上表现效果不如 PMG 方法。分析原因可知，由于 PMG 方法是渐进式学习多粒度图像分类特征，通过控制超参数 n 可以在每次迭代过程中学习到不同粒度下的局部特征，最终再对这些特征进行融合，提高了分类精度。但是，PMG 方法是一种用时间换精度的方法，时间开销比较大。

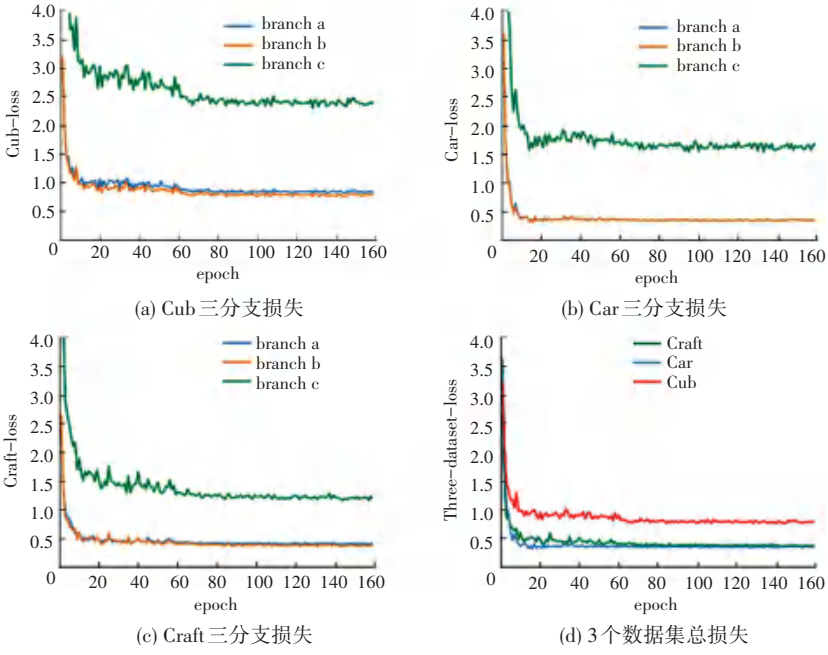
在 CUB-200-2011、FGVC-Aircraft 和 Stanford Cars 数据集上各个分支的损失变化情况和总损失率的变化趋势如图 5 所示。

由图 5 可见，随着迭代次数的增多，各个分支的损失率逐渐收敛。然而，不论哪个数据集，分支 c 的损失率是最大的，因为分支 c 破坏了前景目标的全局结构，在网络训练过程中致力于找到局部判别性特征，因此在总损失函数中，对分支 c 施加权重 γ，设置为 0.5。

表 5　不同弱监督下细粒度图像分类方法实验对比

Table 5　Experimental comparison of different weakly supervised fine-grained image classification methods

方法	Backbone	CUB-200-2011	FGVC-Aircraft	Stanford Cars
Cross-X[19]	ResNet-50	87.7	92.6	94.6
DFL-CNN[20]	ResNet-50	87.4	91.7	93.1
MC-Loss	ResNet-50	87.3	92.6	93.7
NTS-Net	ResNet-50	87.5	91.4	93.9
DCL	ResNet-50	87.8	93.0	94.5
PMG	ResNet-50	**89.6**	93.4	**95.1**
本文方法	ResNet-50	88.1	**93.8**	94.9



(a) Cub 三分支损失　(b) Car 三分支损失

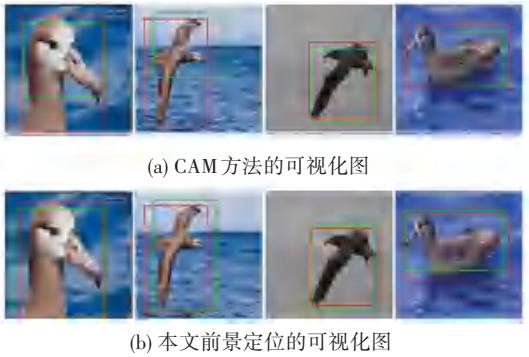
(c) Craft 三分支损失　(d) 3 个数据集总损失

图 5　数据集的三分支损失变化情况和总损失率的变化情况

Fig. 5　The change of three-branch loss and total loss of three datasets

3.4　可视化

对 CAM 方法和本文提出的前景定位方法进行可视化分析,其结果如图 6 所示。图 6 中红色矩形代表真值框,绿色矩形代表网络学习到的边界框。



(a) CAM 方法的可视化图

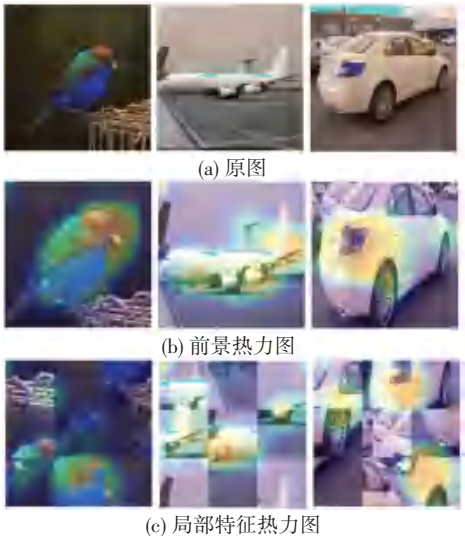
(b) 本文前景定位的可视化图

图 6　目标定位可视化效果图

Fig. 6　Comparison of object localization visual effect

从图 6 的对比可以看出,本文方法的性能优于 CAM 方法,可以检测出较为准确的边界框,消除更多的背景噪声信息,在剔除背景噪声信息的同时不会破坏前景目标的完整性,也能获得更精准的前景目标特征,进一步提高模型的表达能力。

输入图像在本文方法处理后的可视化如图 7 所示。图 7(b)放大前景目标,消除了背景噪音干扰,强化特征提取;图 7(c)破坏前景目标的整体结构,使网络关注了更多的局部判别性特征。



(a) 原图

(b) 前景热力图

(c) 局部特征热力图

图 7　可视化图

Fig. 7　Effect of visualization

4　结束语

本研究提出了一种细粒度图像分类方法,通过结合前景定位和局部判别性特征提取,有效减少背景干扰并得到多样的目标特征。本文做了充分实验验证了跟踪器的性能,后续会从轻量化的模型着手,提高效率。

参考文献

[1] HUANG Shaoli, XU Zhe, TAO Dacheng, et al. Part-stacked

CNN for fine-grained visual categorization[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway,NJ:IEEE, 2016: 1173-1182.

[2] LIN Di, SHEN Xiaoyong, LU Cewu, et al. Deep LAC: Deep localization, alignment and classification for fine-grained recognition [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ:IEEE, 2015: 1666-1674.

[3] XIAO Tianjun, XU Yichong, YANG Kuiyuan, et al. The application of two-level attention models in deep convolutional neural network for fine-grained image classification [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway,NJ:IEEE, 2015: 842-850.

[4] ZHENG Heliang, FU Jianlong, MEI Tao, et al. Learning multi-attention convolutional neural network for fine-grained image recognition [C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Piscataway,NJ:IEEE, 2017: 5209-5217.

[5] ZHOU Bolin, KHOSLA A, LAPEDRIZA A, et al. Learning deep features for discriminative localization [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway,NJ:IEEE, 2016: 2921-2929.

[6] SINGH K K, LEE Y J. Hide-and-Seek: Forcing a network to be meticulous for weakly-supervised object and action localization [C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ:IEEE, 2017: 3544-3553.

[7] YUN S, HAN D, CHUN S, et al. CutMix: Regularization strategy to train strong classifiers with localizable features[C]// Proceedings of the IEEE International Conference on Computer Vision (ICCV). Piscataway,NJ:IEEE, 2019: 6022-6031.

[8] ZHANG Xiaolin, WEI Yunchao, FENG Jiashi, et al. Adversarial complementary learning for weakly supervised object localization [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ:IEEE, 2018: 1325-1334.

[9] CHOE J, LEE S, SHIM H. Attention-based dropout layer for weakly supervised single object localization and semantic segmentation [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ:IEEE, 2019: 2214-2223.

[10] WEI Jun, WANG Qin, LI Zhen, et al. Shallow feature matters for weakly supervised object localization[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway,NJ:IEEE, 2021: 5989-5997.

[11] DING Yao, ZHOU Yanzhao, ZHU Yi, et al. Selective sparse sampling for fine-grained image recognition[C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Piscataway,NJ:IEEE, 2019: 6598-6607.

[12] YANG Ze, LUO Tiange, WANG Dong, et al. Learning to navigate for fine-grained classification[M]//Computer Vision-ECCV2018. Lecture Notes in Computer Science. Cham:Springer, 2018: 420-435.

[13] CHANG Dongliang, DING Yifeng, XIE Jiyang, et al. The devil is in the channels: Mutual-channel loss for fine-grained image classification[J]. IEEE Transactions on Image Processing, 2020 (29): 4683-4695.

[14] CHEN Yue, BAI Yalong, ZHANG Wei, et al. Destruction and construction learning for fine-grained image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2019: 5152-5161.

[15] DU Ruoyi, CHANG Dongliang, BHUNIA A K, et al. Fine-grained visual classification via progressive multi-granularityt raining of Jigsaw patches[M]//VEDALDI A, BISCHOF H, BROX T, et al. Computer Vision. ECCV 2020. Lecture Notes in Computer Science(). Cham: Springer, 2020,12365:153-168.

[16] KRAUSE J, STARK M, DENG J, et al. 3D Object Representations for Fine-grained Categorization[C]//2013 IEEE International Conference on Computer Vision Workshops (ICCVW). Piscataway,NJ:IEEE, 2013: 554-561.

[17] TAN Min, WANG Guijun, ZHOU Jian, et al. Fine-grained classification via hierarchical bilinear pooling with aggregated slack mask[J]. IEEE Access, 2019, 7: 117944-117953.

[18] WEI Xiushen, LUO Jianhao, WU Jianxin, et al. Selective convolutional descriptor aggregation for fine-grained image retrieval[J]. IEEE Transactions on Image Processing, 2017, 26 (6): 2868-2881.

[19] LUO Wei, YANG Xitong, MO Xianjie, et al. Cross-X learning for fine-grained visual categorization[C]// Proceedings of the IEEE International Conference on Computer Vision (ICCV). Piscataway,NJ:IEEE, 2019: 8241-8250.

[20] SHI Xiruo, XU Liutong, WANG Pengfei, et al. Beyond the attention: Distinguish the discriminative and confusable features for fine-grained image classification[C]// Proceedings of the 28th ACM International Conference on Multimedia. New York:ACM, 2020: 601-609.